Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2017

Genome assembly and annotation ofArabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology

Briskine, Roman V ; Paape, Timothy ; Shimizu-Inatsugi, Rie ; Nishiyama, Tomoaki ; Akama, Satoru ; Sese, Jun ; Shimizu, Kentaro K

Abstract: The self-incompatible species Arabidopsis halleri is a close relative of the self-compatible model plant Arabidopsis thaliana. The broad European and Asian distribution and heavy metal hyperaccumulation ability makes A. halleri a useful model for ecological genomics studies.We used long-insert mate-pair libraries to improve the genome assembly of the A. halleri ssp.gemmifera Tada mine genotype (W302) collected from a site with high contamination by heavy metals in Japan. After five rounds of forced selfing, heterozygosity was reduced to 0.04%, which facilitated subsequent genome assembly. Our assembly now covers 196 Mb or 78% of the estimated genome size and achieved scaffold N50 length of 712 kb. To validate assembly and annotation, we used synteny of A. halleri Tada mine with a previously published high quality reference assembly of a closely related species, Arabidopsis lyrata. Further validation of the assembly quality comes from synteny and phylogenetic analysis of the HEAVY METAL ATPASE4 (HMA4) and METAL TOLERANCE PROTEIN1 (MTP1) regions using published sequences from European A. halleri for comparison. Three tandemly duplicated copies of HMA4, key gene involved in cadmium and zinc hyperaccumulation, were assembled on a single scaffold.The assembly will enhance the genome-wide studies of A. halleri as well as the allopolyploid Arabidopsis kamchatica derived from A. lyrata and A. halleri.

DOI: https://doi.org/10.1111/1755-0998.12604

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-126774
Journal Article
Accepted Version

Originally published at:
Briskine, Roman V; Paape, Timothy; Shimizu-Inatsugi, Rie; Nishiyama, Tomoaki; Akama, Satoru; Sese, Jun; Shimizu, Kentaro K (2017). Genome assembly and annotation ofArabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. Molecular Ecology Notes, 17(5):1025-1036.
DOI: https://doi.org/10.1111/1755-0998.12604

Accepted Article

Genome assembly and annotation of *Arabidopsis halleri*, a model for heavy metal hyperaccumulation and evolutionary ecology

Roman V. Briskine[1], Timothy Paape[1], Rie Shimizu-Inatsugi[1], Tomoaki Nishiyama[2], Satoru Akama[3], Jun Sese[3], and Kentaro K. Shimizu[1,4]

[1]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, Zurich, CH-8057, Switzerland
[2]Advanced Science Research Center, Kanazawa University, 13-1 Takara-machi, Kanazawa, 920-0934, Japan
[3]Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi Koto-ku, Tokyo, 135-0064, Japan
[4]Kihara Institute for Biological Research, Yokohama City University, 641-12 Maioka, Totsuka-ward, Yokohama, 244-0813, Japan

Keywords:
Arabidopsis halleri, *de novo* assembly, functional annotation, heavy metal hyperaccumulator, Tada mine

Corresponding author:
Kentaro Shimizu
Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, CH-8057, Switzerland
email: kentaro.shimizu@ieu.uzh.ch
fax: +41 44 635 68 21

Running title: Genome assembly of *Arabidopsis halleri*

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi:

10.1111/1755-0998.12604

This article is protected by copyright. All rights reserved.

Accepted Article

Abstract

The self-incompatible species *Arabidopsis halleri* is a close relative of the self-compatible model plant *Arabidopsis thaliana*. The broad European and Asian distribution and heavy metal hyperaccumulation ability makes *A. halleri* a useful model for ecological genomics studies. We used long-insert mate-pair libraries to improve the genome assembly of the *A. halleri* ssp. *gemmifera* Tada mine genotype (W302) collected from a site with high contamination by heavy metals in Japan. After five rounds of forced selfing, heterozygosity was reduced to 0.04%, which facilitated subsequent genome assembly. Our assembly now covers 196 Mb or 78% of the estimated genome size and achieved scaffold N50 length of 712 kb. To validate assembly and annotation, we used synteny of *A. halleri* Tada mine with a previously published high quality reference assembly of a closely related species, *Arabidopsis lyrata*. Further validation of the assembly quality comes from synteny and phylogenetic analysis of the *HEAVY METAL ATPASE4* (*HMA4*) and *METAL TOLERANCE PROTEIN1* (*MTP1*) regions using published sequences from European *A. halleri* for comparison. Three tandemly duplicated copies of *HMA4*, key gene involved in cadmium and zinc hyperaccumulation, were assembled on a single scaffold. The assembly will enhance the genome-wide studies of *A. halleri* as well as the allopolyploid *Arabidopsis kamchatica* derived from *A. lyrata* and *A. halleri*.

This article is protected by copyright. All rights reserved.

Accepted Article

Introduction

Ecological genomics studies in plant species can now be extended to close relatives of model and to non-model species by improvements in *de novo* assembly and gene annotation methods (Slotte *et al.* 2013; Lobréaux *et al.* 2014; Liu *et al.* 2014). *Arabidopsis halleri* (2*n* = 16) is a close relative of the model plant *A. thaliana* (2*n* = 10) (Novikova *et al.* 2016) and is itself becoming a model for ecological and evolutionary genetics studies. Available genetic tools include transgenic techniques using tissue culture (Hanikenne *et al.* 2008), complementation test using viable F1 hybrid plants with *A. thaliana* mutants (Shimizu 2002), and QTL maps using *Arabidopsis lyrata* (2*n* = 16) (Willems *et al.* 2007). Self-incompatibility in *A. halleri* maintains high genetic diversity in the species, on average an order of magnitude over *A. thaliana* (Castric *et al.* 2008; Roux *et al.* 2011). A high quality genome assembly of *A. lyrata* was recently published (Hu *et al.* 2011) and is the second well assembled *Arabidopsis* genome along with *A. thaliana*. The *A. lyrata* assembly was generated from a self-compatible accession, but no self-compatible population of *A. halleri* has been reported.

A. halleri has attracted the study of speciation and ecological adaptation (Ramos-Onsins *et al.* 2004). The split of *A. halleri* and *A. lyrata* is estimated to coincide with the tandem duplication of *HEAVY METAL ATPASE4* (*HMA4*), suggesting it was an ecological speciation event where *A. halleri* evolved heavy metal tolerance (Roux *et al.* 2011). Although gene flow between the two species is limited, occasional introgression events including *S*-haplogroups were detected (Ramos-Onsins *et al.* 2004; Castric *et al.* 2008). The self-compatible allotetraploid species *Arabidopsis kamchatica* (Tsuchimatsu *et al.* 2012) is derived from the hybridization of *A. halleri* and *A. lyrata* and has a broad climatic niche compared with parental species (Hoffmann 2005; Shimizu *et al.* 2005; Shimizu-Inatsugi *et al.* 2009; Schmickl *et al.* 2010; Novikova *et*

This article is protected by copyright. All rights reserved.

al. 2016). Other molecular ecological and evolutionary studies using *A. halleri* include defense against herbivores by heavy metals (Kazemi-Dinan *et al.* 2014) and by trichomes (Shimizu 2002; Kawagoe *et al.* 2011), local adaptation (Fischer *et al.* 2013; Kubota *et al.* 2015), self-incompatibility and mating systems (Shimizu & Purugganan 2005; Bechsgaard *et al.* 2006; Goubet *et al.* 2012; Durand *et al.* 2014), and gene expression in natural environments (*in natura*) (Aikawa *et al.* 2010; Shimizu *et al.* 2011; Kudoh 2015).

A major area of research in *A. halleri* has focused on the study of heavy metal tolerance and hyperaccumulation (Bert *et al.* 2000; Chiang *et al.* 2006; Hanikenne *et al.* 2008; Krämer 2010). Heavy metal hyperaccumulation is a constitutive phenotype in *A. halleri* and all tested genotypes are able to accumulate high levels of cadmium and zinc in leaves (Kubota & Takenaka 2003; Pauwels *et al.* 2006; Talke *et al.* 2006; Chiang *et al.* 2006). QTL studies using crosses between non-hyperaccumulator *A. lyrata* and *A. halleri* showed that the two major loci that co-segregate with cadmium and zinc tolerance are the *HEAVY METAL ATPASE4* (*HMA4*) and the cation diffusion facilitator (CDF) protein *METAL TOLERANCE PROTEIN1* (*MTP1* also called *ATCDF1* or *ZAT1*) (Willems *et al.* 2007; Courbot *et al.* 2007). The sequencing of a BAC assembly in the *A. halleri* Langelsheim genotype, which includes three paralogous *HMA4* copies (Hanikenne *et al.* 2008), has led to the estimation of a hard selective sweep at *HMA4* in *A. halleri* based on surrounding genetic diversity (Hanikenne *et al.* 2013). BAC sequences of the *A. halleri* genotype from the Auby mine site showed that *MTP1* is duplicated with up to five copies in this population (Shahzad *et al.* 2010). Overall, conserved sequence diversity of duplicated copies of *HMA4* and *MTP1* along with exceptionally high additive expression likely explains constitutive hyperaccumulation in *A. halleri*.

Among the largest genomic regions of *A. halleri* that have been published are BAC assemblies

This article is protected by copyright. All rights reserved.

Accepted Article

of the *HMA4* region (290 kb) from the Langelsheim accession (Hanikenne *et al.* 2008, 2013), the *MTP1-A* region (110 kb) from the Auby mine accession (Shahzad *et al.* 2010), and the *S*-locus of multiple haplotypes (ranging from 25 kb to 121 kb) (Goubet *et al.* 2012). The first two regions harbor genes involved in heavy metal hyperaccumulation and tolerance. Previously we constructed a medium quality *de novo* assembly of *A. halleri* ssp. *gemmifera* (W302) collected at the Tada mine site in Japan to establish a next-generation read sorting pipeline (HomeoRoq) for distinguishing homeolog origins of RNA-seq reads in synthesized *A. kamchatica* (Akama *et al.* 2014). That assembly also contained several long scaffolds over 100 kb. More recently, another accession of *A. halleri* ssp. *gemmifera* (IBO380) from Mt. Ibuki, Japan, was sequenced for genome wide selection scans for altitudinal adaptation (Kubota *et al.* 2015). While the two reference assemblies reported in Akama et al. (2014) and Kubota et al. (2015) covered ~ 88 and 98% of the *A. halleri* genome respectively, they were comprised mostly of short contigs (N50 length ~ 18 kb and 5 kb respectively). Another study, focusing on local adaptation of five populations in the Swiss Alps, analyzed pooled resequencing data using the genome of *A. thaliana* as a reference due to the lack of a high quality genome of *A. halleri*, which limited their analyses to conserved coding regions (Fischer *et al.* 2013). For such studies to identify candidate single nucleotide polymorphisms (SNPs) under selection, long scaffolds would be valuable for detecting long-range linkage disequilibrium as a result of recent selection. Finally, as sequence data accumulates for *A. halleri* along with phenotypic data, long scaffolds are essential for identifying the genetic architecture of quantitative traits.

We present a long-scaffold genome assembly of *A. halleri* ssp. *gemmifera* (Tada mine) that was constructed using long mate-pair libraries (insert sizes ranging from 2.5 kb to 22 kb) along with the existing short insert paired-end libraries from Akama et al. (2014). The Tada mine site

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

in Japan produced silver and copper for about 1,000 years (Kobata 1968) until its commercial shutdown in 1973 (Tada Silver Mine Historic Site Preservation Association 2007). It served as a resource for the Toyotomi dynasty during the 16th century (Azuchi-Momoyama period) and then experienced the peak of production around 1660's (Edo period). This site is contaminated with both cadmium and zinc (Paape *et al.* 2016) (see Supporting Information for details) and *A. halleri* plants are abundant in this area along with another hyperaccumulator, *Athyrium yokoscense*, which was traditionally used as an indicator species to find ore deposits in Japan (Miyake 1897). While *A. halleri* is obligately outcrossing, we drastically reduced heterozygosity in the Tada mine accession by forced self-fertilization (5 times through bud pollination), making assembly more straightforward with higher homozygosity. We compared the synteny of the Tada mine assembly to the fully assembled *A. thaliana* (The Arabidopsis Genome Initiative 2000) and *A. lyrata* (Hu *et al.* 2011) genomes, and to duplicated *HMA4* and *MTP1* regions from previously published BAC sequences of *A. halleri* (Shahzad *et al.* 2010; Hanikenne *et al.* 2013). The successful assembly of complex gene duplications in our Tada mine reference assembly may facilitate evolutionary studies of duplicated genes that led to a multi-locus adaptive phenotype in *A. halleri*.

Materials and methods

Study species and samples

Arabidopsis halleri (L.) O'Kane & Al-Shehbaz (Basionym *Arabis halleri*) is a diploid species (2n = 16) distributed in Europe and East Asia (O'Kane & Al-Shehbaz 1997; Al-Shehbaz & O'Kane 2002). Populations of *A. halleri* in both European and East Asian regions are distributed across highly variable lowland and alpine areas (Fischer *et al.* 2013; Kubota *et al.* 2015) that experience

This article is protected by copyright. All rights reserved.

extremes in temperature, solar radiation, and precipitation. In addition, *A. halleri* is also found across highly variable soil types and is able to tolerate extreme heavy metal contamination in Europe and Asia (Hanikenne *et al.* 2013). The split of *A. halleri* from *A. thaliana* has long been considered to be about 5 million years ago based on the estimation of evolutionary rate by Koch et al. (2000), while more recent estimation by Ossowski et al. (2010) suggested 13.0 or 17.9 million years ago (reviewed by Shimizu & Tsuchimatsu 2015). Subsequently the speciation of *A. halleri* and *A. lyrata* occurred ~337,000 years ago or 2.5 million years ago depending on assumptions (Castric *et al.* 2008; Roux *et al.* 2011). Allopolyploid origins of *A. kamchatica* by the hybridization of *A. halleri* and *A. lyrata* occurred much more recently (~20 thousand years ago) (Tsuchimatsu *et al.* 2012).

A number of subspecies of *A. halleri* have been proposed based on morphology. Al-Shehbaz & O'Kane recognized three subspecies (ssp. *halleri* and ssp. *ovirensis* in Europe, and ssp. *gemmifera* in East Asia). Kolník and Marhold (2006) added ssp. *tatrica* and ssp. *dacica* from East Europe. We consider East Asian plants as *A. halleri* ssp. *gemmifera* (Matsumura) O'Kane & Al-Shehbaz (also called *Arabis gemmifera* or *Cardamine gemmifera*) (O'Kane & Al-Shehbaz 1997; Hoffmann 2005). We sequenced an accession called "Tada mine" (W302 in our stock number), which had undergone 5 rounds of selfing by bud pollination starting from the original *A. halleri* ssp. *gemmifera* individual collected from a population close to a stream near Tada mine in Inagawa city, Hyogo Prefecture, Japan (N 34.89°, E 135.35°, altitude 140 m). A population genetic study using microsatellite markers by Sato and Kudoh (2014) showed that individuals from the same population belong to a widespread cluster in Kinki area, Japan. Soil samples from this site were measured for 16 elements including the heavy metals cadmium and zinc (Supporting File 1).

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Genome size estimation by flow cytometry

We measured the nuclear DNA content of Tada mine accession by flow cytometry. Flower petals were processed together with internal reference standard Tomato leaf (*Lycopersicon esculentum* cv. 'Stupicke'; assumed genome size of 958 Mb) (Doležel *et al.* 1992) and the genome size was calculated from the relative peak position using CyStain PI absolute P kit (Partec) and CyFlow Space (Partec). The obtained genome size (250 Mb) was slightly larger than our previous estimate of 220 Mb (Akama *et al.* 2014).

Construction of Illumina libraries

The construction of paired-end libraries was previously reported in Akama et al. (2014). DNA samples for the long insert mate-pair libraries were taken from the clones of the W302 individual used to create the short insert libraries. The mate-pair libraries with six insert ranges (22-38 kb, 15-22 kb, 11-15 kb, 7-11 kb, 5.0-7 kb, and 3.0-5.0 kb) were constructed with Illumina Nextera Mate-pair Library Prep kit with modification to construct large insert sizes. The details can be found in the Supporting Information text. The libraries were sequenced on Illumina HiSeq 2500 at the Functional Genomics Centre Zurich (http://www.fgcz.ch).

De novo assembly

The *A. halleri* genome was assembled from all available untrimmed read libraries (Table 1) with ALLPATHS-LG R50599 (Butler *et al.* 2008). Among mate-pair libraries, 22-38 kb insert library was treated as long-jumping library while the others were regarded as jumping libraries. The assembly process included two steps. First, ALLPATHS-LG was executed with default parameters and expected insert sizes. Then, the insert size parameters were changed to the values

This article is protected by copyright. All rights reserved.

calculated in the first step, and ALLPATHS-LG was run again. The assembly job completed in 66 hours using 20 cores on a Linux cluster with the peak memory utilization of 126 GB.

Improving assembly using synteny

Because *A. halleri* and *A. lyrata* diverged recently (Schmickl *et al.* 2010; Roux *et al.* 2011) and each has 8 chromosomes (Al-Shehbaz & O'Kane 2002), we used the previously published *A. lyrata* reference genome (Hu *et al.* 2011) to perform genome-wide synteny analysis (see Supplementary Figure S1 for assembly pipeline). The complete genome, coding sequences, and gene annotation of *A. lyrata* strain MN47 v1.07 were downloaded from the Phytozome v9.0 website (http://phytozome.jgi.doe.gov). Coding sequences of *A. lyrata* were aligned to the *A. halleri* assembly using BLAT v3.5 (Kent 2002) with default parameters except maximum intron size. Because the longest intron in the *A. lyrata* assembly was 44,703 bp, we set the maximum intron size to 50 kb. Hits were filtered, sorted, and merged into syntenic regions using custom Perl scripts (see the Data Accessibility section). We only considered the hits covering at least 85% of the query sequence and accepted the hit from a syntenic gene even when it did not have the highest score for the locus. If an *A. halleri* scaffold contained two neighboring loci that were syntenic in *A. halleri* to two *A. lyrata* regions located on different chromosomes or more than 100 kb apart, the scaffold was split into two parts by removing the sequence of unknown nucleotides. Scaffolds were only split if the sequence of unknown nucleotides at the cut site spanned at least 50 bp, assuming that longer N-stretches would indicate that the support for contig splicing came only from longer-insert libraries with lower quality and without gap-filling alignments. After this correction, the scaffolds were sorted by length in descending order and named sequentially beginning with scaffold_1. We also used published BAC sequences (see

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

phylogeny sections below) for the *HMA4* region from the Langelsheim accession and the *MTP1* region from the Auby accession to determine synteny for loci containing known duplications in other *A. halleri* accessions. Because these BAC sequences are among the longest published scaffolds in *A. halleri*, the synteny analysis also serves as validation of our assembly for complex regions.

Heterozygosity estimation

To obtain an estimate of genome-wide heterozygosity, we aligned all reads from 200 bp and 500 bp insert libraries against the assembly using BWA v0.7.2 (Li & Durbin 2009) and called variants using HaplotypeCaller from GATK package v3.4-0 (McKenna *et al.* 2010) following established best practices (DePristo *et al.* 2011; Van der Auwera *et al.* 2013). Low quality variants and variants in known repetitive elements were discarded (see Supporting Information for more details). The number of the remaining variants was divided by the total count of non-missing bases with non-zero coverage in the assembly to estimate the heterozygosity level.

Annotation

To annotate the genome of *A. halleri*, we integrated RNA-seq data from leaves and roots (Paape *et al.* 2016) with the AUGUSTUS gene prediction program (Stanke *et al.* 2006; AUGUSTUS Development Team 2014) (see Supplementary Figure S2 for annotation pipeline). Unstranded paired-end 100 bp reads from the *A. halleri* W302_L4 (leaf) and W302_R1 (root) libraries were individually aligned to the *A. halleri* W302 reference genome using STAR v2.4.0i (Dobin *et al.* 2013) with non-default parameters. (For the complete list of utilities and the detailed description of parameters used at each step, see the readme file in the online code repository

This article is protected by copyright. All rights reserved.

Accepted Article

at https://gitlab.com/rbrisk/ahalassembly.) Compared to TopHat v2.0.13 (Trapnell *et al.* 2009) ran with modified parameters, STAR yielded more unique alignments ultimately resulting in a higher number of hints for AUGUSTUS (Supplementary Table S1). Intron hints were extracted from the alignment and merged with repetitive element (*nonexonpart*) hints derived from the RepeatMasker v4.0.5 (Smit *et al.* 1996) output. The merged hints were used for the preliminary AUGUSTUS v3.0.3 run. Introns were extracted from the output and used to generate exon-exon junction database. The original reads were aligned against exon-exon junction sequences using bowtie2 v2.2.4 (Langmead & Salzberg 2012) rather than STAR because splice-aware alignment was not necessary in this case. Spliced reads were removed from the STAR output and the remaining reads were merged with reads that aligned to exon-exon junctions. The merged reads were filtered to include only the read pairs with high quality alignments. The filtered alignments were used to generate intron hints for the final AUGUSTUS run. Human readable functional descriptions were added using the AHRD tool and following its documentation (Tomato Genome Consortium 2012). Reciprocal best BLAST hits were calculated by aligning all coding sequences (the longest transcript per gene) corresponding to one annotation version against all coding sequences corresponding to another annotation version (see Supplementary Table S3 for the list of annotations) both ways using NCBI BLAST+ v2.2.29 and comparing the scores for hits longer than 200 bp.

Phylogeny of *HMA4* duplications

Previously published BAC sequences (GenBank accessions: EU382072.1 and EU382073.1) covering the tandemly duplicated *HMA4* gene in the Langelsheim accession of *A. halleri* ssp. *halleri* (Hanikenne *et al.* 2008, 2013) were used to assess synteny within the scaffold containing

This article is protected by copyright. All rights reserved.

the *HMA4* region in our *A. halleri* Tada mine assembly. Because the two Langelsheim BAC sequences overlap, they were spliced together using the minimus2 application from the Amos v3.1.0 package (Sommer *et al.* 2007). The overlapping BAC sequence assembly is ~290 kb in length and contains three copies of *HMA4* coding sequences and promoter regions, plus flanking gene models (Hanikenne *et al.* 2013). We used AUGUSTUS v3.0.3 (Stanke *et al.* 2006) to predict gene models on the Langelsheim BAC sequence and aligned these models with BLAST v2.2.30+ (Camacho *et al.* 2009) against *A. thaliana* coding sequences in order to assign TAIR10 gene IDs.

We extracted coding sequences for each of the three *HMA4* copies on the Langelsheim *A. halleri* BAC assembly, and obtained *A. lyrata* and *A. thaliana HMA4* orthologous coding sequences from GenBank to use as outgroup sequences. We then used the three predicted *HMA4* coding sequences from our Tada mine assembly to construct a phylogeny. Coding sequences were aligned using Geneious v6.06 (Biomatters), and phylogenetic tree construction was performed using MrBayes (Ronquist *et al.* 2012) and PhyML (Guindon *et al.* 2010) with default parameters. For MrBayes, we selected GTR model and ran it for 500,000 generations sampling every 500 generations.

Phylogeny of *MTP1* duplications

The five duplications include *AhMTP1-A1* and *-A2* on the FN428855.1 BAC sequence and *AhMTP1-B*, *-C* and *-D* on the FN386317, FN386316, and FN386315 BAC sequences respectively. The BAC sequences were downloaded from the EMBL database. The longest BAC (FN428855.1) is ~110 kb and contains the tandem duplications *AhMTP1-A1* and *-A2* including several flanking genes, while the others contain only the *AhMTP1-B*, *-C* and *–D* copies and no flanking genes (on

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

average 5 kb each). We used the same approach as with *HMA4* BACs to annotate the genes on the BAC containing *AhMTP1-A1* and *-A2*. This longest BAC was then used to check synteny with our scaffold containing the *A. halleri* Tada mine *AhMTP1-A1* ortholog.

Results and Discussion

Tada mine assembly version 2.2

Long-insert libraries were used in conjunction with reads from three paired-end libraries constructed for the previous draft assembly v1.0 of *A. halleri* Tada mine (Akama *et al.* 2014) in an attempt to improve scaffold length. While version 1.0 covered 88% of the estimated genome size (250 Mb by flow cytometry), the highly fragmented (low N50) assembly limited analyses to gene coding sequences. We constructed six mate-pair libraries with insert sizes ranging from 3 kb to 38 kb and sequenced them to obtain more than 700 million additional reads (Table 1). A large proportion of the reads were either duplicates or had low complexity and were discarded by the assembler (see '% Used' column in Table 1). The effective coverage (i.e. the coverage of the kept reads) for the mate-pair libraries was over 100x and the total effective coverage encompassing three paired-end and the six mate-pair libraries was almost 200x (Table 1).

The Tada mine v2.2 assembly has a smaller total size compared to v1.0 (196 Mb and 221 Mb respectively; Table 2). The smaller size of v2.2 is due to the contig size filtering automatically performed by the assembler. The shortest scaffold length in v1.0 was 100 bp while it was 932 bp in v2.2. When we removed all scaffolds shorter than 932 bp from v1.0, its total length decreased to 180 Mb. Likewise, the percentage of missing nucleotides is lower in v1.0 because longer scaffolds often represent concatenation of contigs with long stretches of Ns inserted to

This article is protected by copyright. All rights reserved.

preserve expected distance between the contigs. Meanwhile, very short sequences in the previous assembly represented individual contigs without missing data.

Using flow cytometry, we estimated the genome size of Tada mine to be 250 Mb, which is slightly smaller than 255 Mb (Johnston *et al.* 2005) and 279 Mb (Wolf *et al.* 2014) previously reported for other individuals of the species. Thus, the v2.2 assembly covers 78% of the genome or 67% if the missing data is excluded. Other important statistics indicate large improvements in v2.2 over v1.0 (Table 2). In particular, the scaffold N50 and NG50 lengths (Earl *et al.* 2011) increased from 18 kb and 15 kb to 712 kb and 489 kb respectively.

We verified the quality of the assembly using previous Sanger sequencing data of the *S*-locus region of the same accession (Tsuchimatsu *et al.* 2010). The *S*-locus is known to be a difficult region for assembly due to numerous repetitive sequences surrounding self-incompatibility specificity genes (e.g. *SCR/SP11* and *SRK*), and most of the previous studies used BAC library sequencing rather than whole genome assembly (Goubet *et al.* 2012). Despite this difficulty, using alignment by BLAT (Kent 2002), scaffold_461 showed a perfect match and coverage of 1,530 bp encompassing entire exons and an intron of *SCR* haplogroup A, and a perfect match to about half of *SRK* (1,952 bp). The other side of the *SRK* sequence (1,458 bp) had a perfect match to scaffold_113, although 14 bp of the 3424 bps of *SRK* were not found in our assembly. We found no mismatches between Sanger data (about 5 kb) and the assembly, which suggests the high accuracy of the assembly.

Tandemly duplicated regions are also difficult to assemble from short reads necessitating BAC sequencing instead. For example, BAC sequencing showed that *A. halleri* Langelsheim accession had three tandemly duplicated *HMA4* copies. In the Tada mine v1.0 assembly, the *HMA4* locus and surrounding genes were scattered across multiple scaffolds and unassembled

This article is protected by copyright. All rights reserved.

Accepted Article

contigs with only one complete *HMA4* copy present. In version 2.2 of the assembly, the entire region was captured on a single scaffold containing all three tandemly duplicated *HMA4* copies (see below), supporting the high quality of the assembly.

Generally, high heterozygosity has a negative effect on assembly quality (Schatz *et al.* 2012). Because we sequenced a genotype that experienced five rounds of self-fertilization and each selfing is expected to reduce heterozygosity into about half, we should have reduced the heterozygosity into about 1/32. In agreement with this expectation, the heterozygosity that was estimated using the assembly and the genome-wide data of *A. halleri* Tada mine genome is very low (0.0402%). We suggest that this high homozygosity contributed to high quality of the assembly.

Annotation

Using the hints derived from RepeatMasker (Smit *et al.* 1996) output and RNA-seq data, AUGUSTUS (Stanke *et al.* 2006) identified 34,553 putative transcripts corresponding to 32,553 loci (Supplementary Table S2). The number of genes is comparable to *A. lyrata*, for which 32,670 genes were predicted but alternative transcripts were not reported by Hu et al. (2011). The gene number is, however, higher than 28,775 genes in *A. thaliana* TAIR10 annotation (The Arabidopsis Genome Initiative 2000). This can be explained in part by larger genome size of *A. halleri* compared to *A. thaliana* (250 Mb in 8 chromosomes vs 125 Mb in 5 chromosomes respectively). In addition, our annotation may contain more pseudogenes than the more meticulously curated *A. thaliana* annotation. On the other hand, the total of 35,286 transcripts were reported for *A. thaliana* suggesting that AUGUSTUS did not report some of the alternative transcripts in *A. halleri*. Overall, 25,328 coding sequences could be aligned against *A. thaliana* TAIR10 gene

This article is protected by copyright. All rights reserved.

Accepted Article

models and 21,433 of them were reciprocal best BLAST hits (Supplementary Table S3). The remaining alignable genes could be duplicates of genes that only have a single copy in *A. thaliana* or homologous genes that are less diverged in *A. halleri* than *A. thaliana*. It is also possible that some alleles appear as separate sequences due to misassembly.

With hints based on RNA-seq data from root and leaf tissues of *A. halleri* Tada mine, AUGUSTUS identified more gene models than without the RNA-seq hints (32,553 and 29,628 respectively; Supplementary Table S2). The two annotations have 22,733 genes with identical coding sequences but only 9,098 identical gene models. The rest of the genes with matching coding sequence have different UTRs or include extra transcripts. In addition, using RNA-seq hints resulted in 120 instances of gene model fusion (when a gene model in one version included exons from multiple gene models in another version) and 1,358 instances of gene model splits (when a gene model was split into multiple gene models). The RNA-seq based annotation also yields more reciprocal best hits (21,433 vs. 21,018 respectively) with *A. thaliana* TAIR10 annotation.

Synteny-based adjustments

A. halleri is a close relative of *A. thaliana* and *A. lyrata*, both of which have high quality published genomes (The Arabidopsis Genome Initiative 2000; Hu *et al.* 2011). Unlike *A. thaliana* which has only 5 chromosomes, both *A. halleri* and *A. lyrata* have 8 chromosomes (reviewed in Hunter & Bomblies 2010) and extensive synteny is expected between these two species. Successful QTL mapping (Willems *et al.* 2007) also suggest large-scale synteny. To assess the quality of our assembly, we extracted the coding sequences from the *A. lyrata* reference genome (Hu *et al.* 2011) and aligned them against our assembly.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Overall, we found 1,303 syntenic regions. The number is smaller than the total number of scaffolds in our *A. halleri* assembly because some scaffolds did not have any alignment hits. On the other hand, we discovered 725 cases where two neighboring regions located on the same scaffold were syntenic to loci located either on two different chromosomes or more than 100 kb apart. While some of those instances may represent genuine structural rearrangements, the close relationship between the species suggests that many of them are caused by misassembly. Therefore, we took a conservative approach and split a scaffold whenever two such regions were joined by 50 or more N's. Based on these criteria, we made cuts at 454 sites on 162 scaffolds. To differentiate the updated assembly from the assembler's original output, we denote them as v2.2 (cut) and v2.0 (uncut).

Heavy metals in soil

The metal concentrations in soil samples of the Tada mine site were high to be considered heavy metal contaminated (metalliferous) according to Bert et al. (2002) and exceeded several legislative thresholds (Supporting File 1 and Supporting Information). Accordingly, W302 showed the high level of zinc hyperaccumulation (Paape *et al.* 2016). Both of these data sets underline the functional importance of the heavy metal hyperaccumulation genes such as *HMA4* and *MTP1* for individuals growing at the site.

HMA4 (cadmium and zinc ATPase transporter) region

We compared the Langelsheim *HMA4* BAC assembly (Hanikenne *et al.* 2013) and our assembled *HMA4* scaffold for Tada mine (Figure 1 and Supplementary Table S4); we will call them haplotypes hereafter. With slight differences in physical distances among three duplicated *HMA4*

This article is protected by copyright. All rights reserved.

gene copies, surrounding gene synteny is highly similar with some minor differences in flanking genes (see Supporting Information for details). Our assembly of the *HMA4* flanking region is also consistent with that of *A. thaliana* chromosome 2 (TAIR10) and *A. lyrata* LG3 chromosome 3 (Shahzad *et al.* 2010) with minor rearrangements (Supporting Information).

The phylogeny of the *HMA4* homologs (Figure 2) did not show pairing of syntenic copies (for example, *HMA4-1* of Tada mine and Langelsheim accessions did not cluster), but rather the three copies clustered within *A. halleri* accessions. The clustering of the three tandem duplicates of Tada mine accession was highly supported (1.00 posterior probability in the MrBayes tree in Figure 2, 100 bootstrap support in maximum likelihood analysis.) These copies show greater percent identity within Tada mine *A. halleri* than between the two accessions (Figure 2). Because the presence of three *HMA4* copies is widespread in *A. halleri* samples (Hanikenne *et al.* 2013; Paape *et al.* 2016), the data are consistent with gene conversion events homogenizing the three copies in the Tada mine accession (see Supporting Information for details). To conduct a formal test for gene conversion, additional data from several *A. halleri* accessions would be necessary (Mansai & Innan 2010).

MTP1 (metal tolerance protein 1) region

We also checked for synteny in scaffolds containing *MTP1* orthologs in *A. halleri* Tada mine and an accession of *A. halleri* ssp. *halleri* from Auby mine, which was shown to possess five paralogous copies of *MTP1* located on three separate chromosomes (Shahzad *et al.* 2010). These five duplicated copies of *MTP1* genes were captured on four BAC sequences in the Auby mine accession. Using a BLAST homology search, we detected three rather than five *MTP1* paralogs on three separate scaffolds in our *A. halleri* Tada mine assembly. This is unsurprising considering

This article is protected by copyright. All rights reserved.

Accepted Article

that the study of Shahzad et al. (2010) did not detect five copies in all genotypes of *A. halleri* ssp. *halleri* where some genotypes showed three or four copies rather than five. Dräger et al. (2004) also identified only three *MTP1* copies in the Langelsheim accession.

The longest of the four Auby BAC sequences contains a tandem duplication of *MTP1* (*AhMTP1-A1* and *–A2*) where both copies are nearly identical at the nucleotide level (Shahzad *et al.* 2010). We can infer that the *MTP1-A* copy in *A. halleri* is the ancestral copy of the other duplicated copies as this region is syntenic with both *A. thaliana* and *A. lyrata* (Figure 3; Supplementary Table S5), each containing only a single *MTP1* copy. In addition to synteny with corresponding *A. thaliana* and *A. lyrata* chromosomal regions containing *MTP1*, the order of genes surrounding *MTP1-A* on the Auby BAC assembly and Tada mine scaffold_22 (*g09643*) is also highly similar with the same adjacent gene models as in *A. thaliana* and *A. lyrata*. Unlike the Auby mine *MTP1-A* region, the Tada mine scaffold_22 does not contain a tandem duplication of *MTP1-A*, which indicates an independent gene duplication in the Auby lineage (Figure 4). Despite marginal support for some nodes (e.g. 0.65 at the node connecting Auby and Tada mine *MTP1-A* copies) in the *MTP1* phylogeny, it is important to note that each *MTP1* copy in Tada mine clusters with the corresponding Auby mine *MTP1* copy. This is supported by synteny of the surrounding genes in our scaffolds containing each *MTP1* copy with the *A. lyrata* genome (Table S5; also see reciprocal best BLAST hits deposited to Dryad) and by QTL maps generated by crosses of *A. halleri* and *A. lyrata* (Willems *et al.* 2007; Shahzad *et al.* 2010). The *A. lyrata* assembly analysis, along with the markers used in Shahzad et al. (2010), indicates that the three Tada mine genes *g10163* (on scaffold_22), *g28207* (scaffold_154), and *g18715* (scaffold_61) are syntenic with *MTP1-A*, *MTP1-B*, and *MTP1-C* respectively in Auby mine *A. halleri*.

This article is protected by copyright. All rights reserved.

Accepted Article

Accepted Article

Conclusion

The recent interest in developing other model Brassicaceae systems has been facilitated by advances in genome assembly and DNA polymorphism datasets. This family possesses many species with unique global distributions, mating systems, life histories, and adaptations. Within the genus *Arabidopsis*, *A. thaliana*, *A. lyrata*, *A. halleri*, and *A. kamchatica* provide opportunities to study ecological genomics under different temperature regimes, along latitudinal and longitudinal clines, variable altitudes, and variable soil types. The long-scaffold assembly of the *A. halleri* Tada mine accession can now be used to identify long-range patterns of polymorphism and diversity, and further genotype – phenotype association studies where knowledge of the genetic architecture of complex phenotypes such as flowering time and heavy metal tolerance is needed.

Acknowledgments

We thank the Functional Genomic Center Zurich for technical support; Hiroshi Kudoh, Kiyotaka Okada, Nobuyoshi Nakajima, and Heidi Lischer for discussion and logistic support; Eiji Izawa and Mika Aoki for the information on the history of Tada mine. The study was supported by Swiss National Science Foundation to KKS, the University Research Priority Program of Evolution in Action of the University of Zurich to KKS and RSI, MEXT KAKENHI Grant Number 16H06469, 26113709 and Young Investigator Award of Human Frontier Science Program to KKS and JS, European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no GA-2010-267243 – PLANT FELLOWS to RVB and TP, Marie-Heim Hoegtlin grant by Swiss National Science Foundation to RSI.

This article is protected by copyright. All rights reserved.

Accepted Article

Data Accessibility

The raw Illumina sequences of the mate-pair libraries were deposited to the DNA Data Bank of Japan (DDBJ) under the project number PRJDB4382 and were given the accession numbers DRX045069 – DRX045074. Paired-end libraries were previously deposited to DDBJ by Akama et al. (2014) under accession numbers DRX012199 –DRX012201.

Assembled sequence was deposited to the European Nucleotide Archive (ENA) and was given accession number ERZ270560. Functional annotation, lists of reciprocal best BLAST hits between *A. halleri* and individually *A. thaliana* and *A. lyrata*, and the list of transposable elements were deposited to DryAd (doi:10.5061/dryad.gn4hh).

Workflow documentation and all custom scripts used in the project have been deposited to GitLab (https://gitlab.com/rbrisk/ahalassembly).

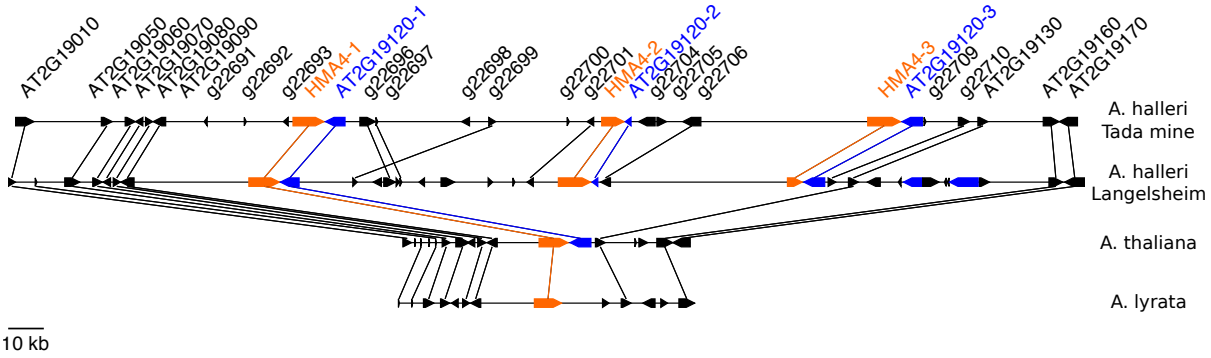
Author Contributions

RVB, TP, RSI, JS and KKS designed research; RSI and TN performed experiments; RVB, TP, SA analyzed the data with inputs from all others; RVB, TP, RSI, and KKS wrote the paper with inputs from all others.

This article is protected by copyright. All rights reserved.

Figures

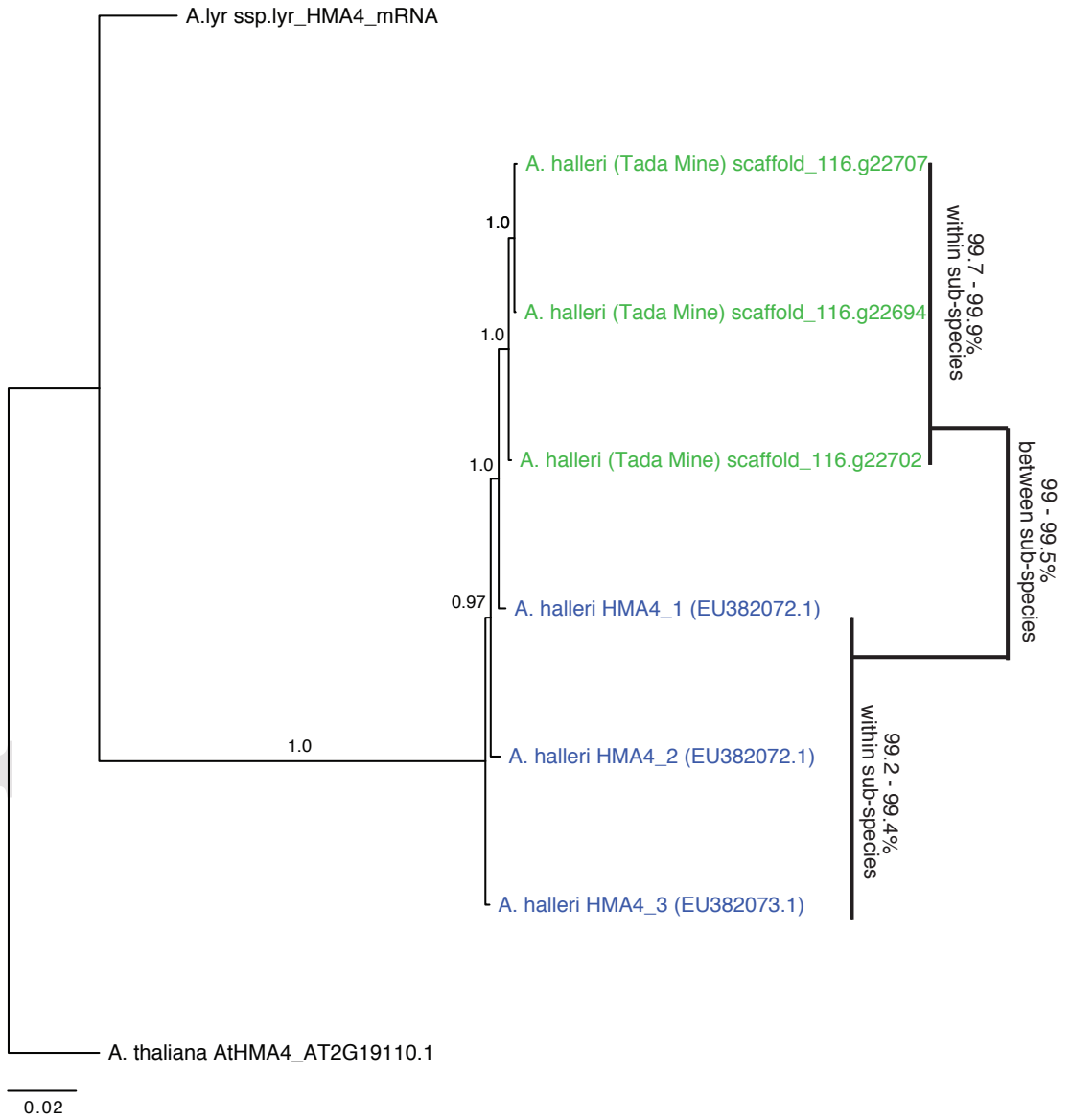
Figure 1: Synteny among *HMA4* regions of *A. halleri* Tada mine, *A. halleri* Langelsheim (Hanikenne *et al.* 2008), *A. thaliana* (TAIR10), and *A. lyrata* MN47 (Hu *et al.* 2011). Genes connections are based on all-vs-all BLAST hits among all coding sequences in the region. *HMA4* copies are highlighted in orange. We connected HMA4 gene copies according to their order. The first copy, *HMA4-1*, is connected in this figure to *AT2G19110.1* in *A. thaliana*. The distance between *HMA4-1* and *HMA4-2* is about 73 kb in both of the *A. halleri* accessions. However, the distance between *HMA4-2* and *HMA4-3* is 64 kb and 52 kb in Tada mine and Langelsheim respectively. *AT2G19120* copies are displayed in blue. Compared to the Tada mine accession, Langelsheim possesses two additional copies of the gene.



This article is protected by copyright. All rights reserved.

Accepted Article

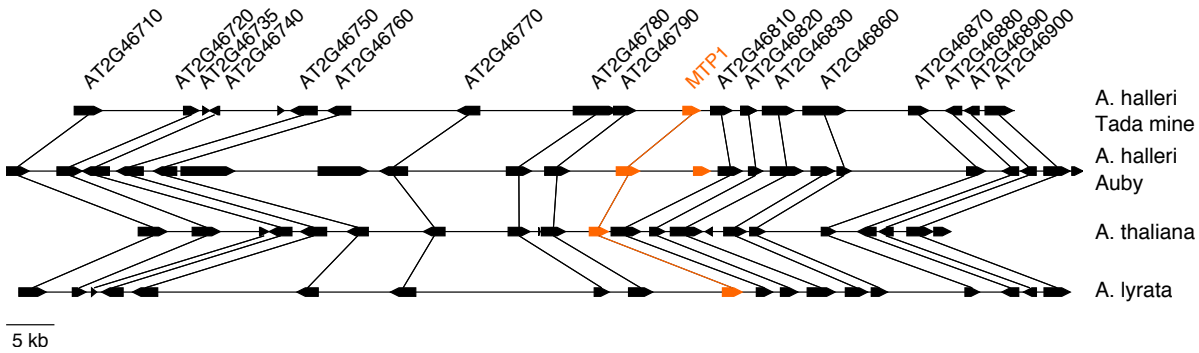
Figure 2: Phylogenetic relationships among triplicated paralogous *HMA4* genes (coding sequence only) in *A. halleri* Tada mine (green, scaffold_116) and the published *A. halleri HMA4* BAC sequences from the Langelsheim accession (blue, EU382072.1 and EU382073.1 (Hanikenne *et al.* 2008, 2013)). The tree was rooted using sequences of *A. lyrata* ssp. *lyrata* mRNA (DQ221101.1) and *A. thaliana* coding sequence (TAIR10: *AT2G19110.1*). Branch labels indicated Bayesian posterior probability scores as reported by MrBayes (Ronquist *et al.* 2012).



This article is protected by copyright. All rights reserved.

Accepted Article

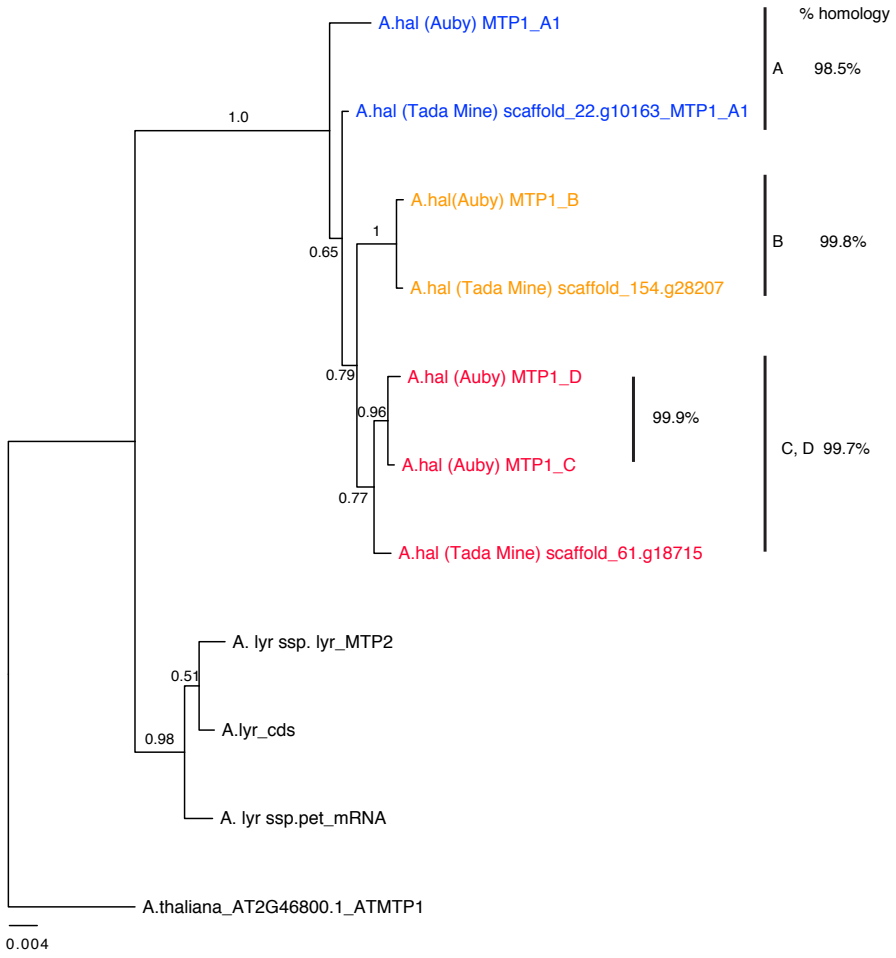
Figure 3: Synteny among *MTP1-A* regions of *A. halleri* Tada mine, *A. halleri* Auby (Shahzad *et al.* 2010), *A. thaliana* (TAIR10), and *A. lyrata* (Hu *et al.* 2011). Gene connections are based on all-vs-all BLAST hits among all coding sequences in the region. *MTP1* is highlighted in orange and is orthologous to *AT2G46800.1* in *A. thaliana*. *A. halleri* Auby contains two copies of the gene.



This article is protected by copyright. All rights reserved.

Accepted Article

Figure 4: Phylogenetic relationships between *A. halleri* Tada mine assembly and *A. halleri* Auby *MTP1* orthologs and paralogs. Colors correspond to the *MTP1* copies on linkage groups and relative to syntenic *A. thaliana* regions in Shahzad et al. (2010). Outgroup sequences are *A. lyrata* ssp. *petraea* (AJ704807.1), *A. lyrata* ssp. *lyrata* (XM_002880219.1), *A. lyrata* mRNA (AY483147.1), and *A. thaliana* (AT2G46800.1).



This article is protected by copyright. All rights reserved.

Accepted Article

Tables

Table 1: **Libraries used for assembly.** Coverage values (Cov) are based on the expected nuclear genome size of 250 Mb. The assembler discarded a large number of reads either due to duplication or low complexity. The percentage of the kept reads is reported in the '% Used' column. The effective coverage column (Eff Cov) indicates the coverage from the reads kept by the assembler.

Type	Insert Size	Reads	Cov	% Used	Eff Cov
Paired-End	200 bp	140,506,146	56.8	70.4	40.0
Paired-End	500 bp	128,033,686	51.7	70.8	36.6
Paired-End	800 bp	39,614,066	16.0	71.7	11.5
Total Paired-End		**308,153,898**	**124.5**		**88.1**
Mate-Pair	3-5 kb	171,385,818	69.2	52.9	36.6
Mate-Pair	5-7 kb	184,082,138	74.4	51.4	38.2
Mate-Pair	7-11 kb	172,373,480	69.6	16.5	11.5
Mate-Pair	11-15 kb	84,648,858	34.2	16.3	5.6
Mate-Pair	15-22 kb	85,893,936	34.7	46.7	16.2
Mate-Pair	22-38 kb	14,873,342	6.0	6.4	0.4
Total Mate-Pair		**713,257,572**	**288.2**		**108.5**
Total		**1,021,411,470**	**412.7**		**196.6**

This article is protected by copyright. All rights reserved.

Accepted Article

Table 2: ***De novo* assembly statistics by version.** Version 1.0 was published by Akama et al. (2014). Version 2.0 is the final output from the assembler. Version 2.2 contains version 2.0 scaffolds that were split based on the lack of synteny with *A. lyrata* genome. Scaffold N50 (NG50) count and length denote the minimum number and the shortest of the scaffolds needed to cover 50% of the assembly (expected genome) length respectively (Earl *et al.* 2011). NG50 values are based on the expected nuclear genome size of 250 Mb as determined by flow cytometry.

Assembly	v1.0	v2.0	v2.2
Total, bp	221,139,660	197,184,962	196,243,198
Missing %	11.81	15.21	14.81
Scaffold #	282,453	1,788	2,239
Shortest, bp	100	932	932
Longest, bp	173,717	7,041,476	4,302,264
N50 length, bp	17,686	1,325,478	712,249
N50 count	3,206	44	71
NG50 length, bp	13,752	907,610	489,153
NG50 count	4,133	69	117

This article is protected by copyright. All rights reserved.

Accepted Article

References

Aikawa S, Kobayashi MJ, Satake A, Shimizu KK, Kudoh H (2010) Robust control of the seasonal expression of the *Arabidopsis FLC* gene in a fluctuating environment. *P Natl Acad Sci USA*, **107**, 11632–11637.

Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J (2014) Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid *Arabidopsis*. *Nucleic Acids Res*, **42**, e46–e46.

Al-Shehbaz IA, O'Kane SL (2002) Taxonomy and phylogeny of *Arabidopsis* (Brassicaceae). *The Arabidopsis Book / American Society of Plant Biologists*, **1**, e0001.

AUGUSTUS Development Team (2014) Incorporating RNAseq data into AUGUSTUS with TopHat. URL http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=IncorporatingRNAseq.Tophat.

Bechsgaard JS, Castric V, Charlesworth D, Vekemans X, Schierup MH (2006) The transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 myr. *Mol Biol Evol*, **23**, 1741–1750.

Bert V, Bonnin I, Saumitou-Laprade P, De Laguérie P, Petit D (2002) Do *Arabidopsis halleri* from nonmetallicolous populations accumulate zinc and cadmium more effectively than those from metallicolous populations? *New Phytol*, **155**, 47–57.

Bert V, Macnair MR, Laguerie P de, Saumitou-Laprade P, Petit D (2000) Zinc tolerance and accumulation in metallicolous and nonmetallicolous populations of *Arabidopsis halleri* (Brassicaceae). *New Phytol*, **146**, 225–233.

Butler J, MacCallum I, Kleber M *et al.* (2008) ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res*, **18**, 810–820.

Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: Architecture and applications. *BMC Bioinformatics*, **10**, 421–421.

Castric V, Bechsgaard J, Schierup MH, Vekemans X (2008) Repeated adaptive introgression at a gene under multiallelic balancing selection. *PLoS Genet*, **4**, e1000168.

Chiang H-C, Lo J-C, Yeh K-C (2006) Genes associated with heavy metal tolerance and accumulation in Zn/Cd hyperaccumulator *Arabidopsis halleri*: A genomic survey with cDNA microarray. *Environ Sci Technol*, **40**, 6792–6798.

Courbot M, Willems G, Motte P *et al.* (2007) A major quantitative trait locus for cadmium tolerance in *Arabidopsis halleri* colocalizes with *HMA4*, a gene encoding a heavy metal

This article is protected by copyright. All rights reserved.

ATPase. *Plant Physiol*, **144**, 1052–1065.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, **43**, 491–498.

Dobin A, Davis CA, Schlesinger F *et al.* (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Doležel J, Sgorbati S, Lucretti S (1992) Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol Plantarum*, **85**, 625–631.

Dräger DB, Desbrosses-Fonrouge A-G, Krach C *et al.* (2004) Two genes encoding *Arabidopsis halleri* MTP1 metal transport proteins co-segregate with zinc tolerance and account for high *MTP1* transcript levels. *Plant J*, **39**, 425–439.

Durand E, Méheust R, Soucaze M *et al.* (2014) Dominance hierarchy arising from the evolution of a complex small RNA regulatory network. *Science*, **346**, 1200–1205.

Earl D, Bradnam K, St. John J *et al.* (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res*, **21**, 2224–2241.

Fischer MC, Rellstab C, Tedder A *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. *Mol Ecol*, **22**, 5594–5607.

Goubet PM, Bergès H, Bellec A *et al.* (2012) Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in *Arabidopsis*. *PLoS Genet*, **8**, e1002495.

Guindon S, Dufayard J-F, Lefort V *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol*, **59**,

Accepted Article

This article is protected by copyright. All rights reserved.

Accepted Article

307–321.

Hanikenne M, Kroymann J, Trampczynska A *et al.* (2013) Hard selective sweep and ectopic gene conversion in a gene cluster affording environmental adaptation. *PLoS Genet*, **9**, e1003707.

Hanikenne M, Talke IN, Haydon MJ *et al.* (2008) Evolution of metal hyperaccumulation required *cis*-regulatory changes and triplication of *HMA4*. *Nature*, **453**, 391–395.

Hoffmann MH (2005) Evolution of the realized climatic niche in the genus: *Arabidopsis* (Brassicaceae). *Evolution*, **59**, 1425–1436.

Hu TT, Pattyn P, Bakker EG *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet*, **43**, 476–481.

Hunter B, Bomblies K (2010) Progress and promise in using Arabidopsis to study adaptation, divergence, and speciation. *The Arabidopsis Book*, e0138.

Johnston JS, Pepper AE, Hall AE *et al.* (2005) Evolution of genome size in Brassicaceae. *Ann Bot*, **95**, 229–235.

Kawagoe T, Shimizu KK, Kakutani T, Kudoh H (2011) Coexistence of trichome variation in a natural plant population: A combined study using ecological and candidate gene approaches. *PLoS ONE*, **6**, e22184.

Kazemi-Dinan A, Thomaschky S, Stein RJ, Krämer U, Müller C (2014) Zinc and cadmium hyperaccumulation act as deterrents towards specialist herbivores and impede the performance of a generalist herbivore. *New Phytol*, **202**, 628–639.

Kent WJ (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Res*, **12**, 656–664.

Kobata A (1968) *Nihon Kozanshi-no Kenkyu (Study of History of Japanese Mines)*. Iwanami-shoten, Tokyo (in Japanese).

Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol Biol Evol*, **17**, 1483–1498.

Kolník M, Marhold K (2006) Distribution, chromosome numbers and nomenclature conspect of *Arabidopsis halleri* (Brassicaceae) in the Carpathians. *Biologia*, **61**, 41–50.

Krämer U (2010) Metal Hyperaccumulation in Plants. *Annu Rev Plant Biol*, **61**, 517–534.

Kubota H, Takenaka C (2003) Field note: *Arabis gemmifera* is a hyperaccumulator of Cd and Zn. *Int J Phytoremediat*, **5**, 197–201.

Kubota S, Iwasaki T, Hanada K *et al.* (2015) A genome scan for genes underlying microgeographic-scale local adaptation in a wild *Arabidopsis* species. *PLoS Genet*, **11**,

This article is protected by copyright. All rights reserved.

e1005361.

Kudoh H (2015) Molecular phenology in plants: In natura systems biology for the comprehensive understanding of seasonal responses under natural environments. *New Phytol*.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth*, **9**, 357–359.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Liu S, Liu Y, Yang X *et al.* (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun*, **5**.

Lobréaux S, Manel S, Melodelima C (2014) Development of an *Arabis alpina* genomic contig sequence data set and application to single nucleotide polymorphisms discovery. *Mol Ecol Resour*, **14**, 411–418.

Mansai SP, Innan H (2010) The power of the methods for detecting interlocus gene conversion. *Genetics*, **184**, 517–527.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**, 1297–1303.

Miyake K (1897) Hebinonegoza to koushitsutono kankei (Relationship between *Aspelnium yokoscense* Fr. Et Sav. and mineral). *Bot Mag Tokyo*, **11**, 404–406.

Novikova PY, Hohmann N, Nizhynska V *et al.* (2016) Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat Genet*, **advance online publication**.

Ossowski S, Schneeberger K, Lucas-Lledó JI *et al.* (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.

O'Kane SL, Al-Shehbaz IA (1997) A Synopsis of *Arabidopsis* (Brassicaceae). *Novon*, **7**, 323–327.

Paape T, Hatakeyama M, Shimizu-Inatsugi R *et al.* (2016) Conserved but attenuated parental gene expression in allopolyploids: Constitutive zinc hyperaccumulation in the allotetraploid *Arabidopsis kamchatica*. *Mol Biol Evol*.

Pauwels M, Frérot H, Bonnin I, Saumitou-Laprade P (2006) A broad-scale analysis of population differentiation for Zn tolerance in an emerging model species for tolerance study: *Arabidopsis halleri* (Brassicaceae). *Journal of Evolutionary Biology*, **19**, 1838–1850.

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aguadé M (2004) Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*.

This article is protected by copyright. All rights reserved.

Accepted Article

Genetics, **166**, 373–388.

Ronquist F, Teslenko M, Mark P van der *et al.* (2012) MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*, **61**, 539–542.

Roux C, Castric V, Pauwels M *et al.* (2011) Does speciation between *Arabidopsis halleri* and *Arabidopsis lyrata* coincide with major changes in a molecular target of adaptation? *PLoS ONE*, **6**, e26872.

Sato Y, Kawagoe T, Sawada Y, Hirai M, Kudoh H (2014) Frequency-dependent herbivory by a leaf beetle, *Phaedon brassicae*, on hairy and glabrous plants of *Arabidopsis halleri* subsp. *gemmifera*. *Evol Ecol*, **28**, 545–559.

Schatz M, Witkowski J, McCombie WR (2012) Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol*, **13**, 243.

Schmickl R, Jorgensen M, Brysting A, Koch M (2010) The evolutionary history of the *Arabidopsis lyrata* complex: A hybrid in the amphi-Beringian area closes a large distribution gap and builds up a genetic barrier. *BMC Evol Biol*, **10**, 98.

Shahzad Z, Gosti F, Frérot H *et al.* (2010) The five *AhMTP1* zinc transporters undergo different evolutionary fates towards adaptive evolution to zinc tolerance in *Arabidopsis halleri*. *PLoS Genet*, **6**, e1000911.

Shimizu KK (2002) Ecology meets molecular genetics in *Arabidopsis*. *Popul Ecol*, **44**, 0221–0233.

Shimizu KK, Purugganan MD (2005) Evolutionary and Ecological Genomics of Arabidopsis. *Plant Physiol*, **138**, 578–584.

Shimizu KK, Tsuchimatsu T (2015) Evolution of Selfing: Recurrent Patterns in Molecular Adaptation. *Annu. Rev. Ecol. Evol. Syst.*, **46**, 593–622.

Shimizu KK, Fujii S, Marhold K, Watanabe K, Kudoh H (2005) *Arabidopsis kamchatica* (Fisch. ex DC.) K. Shimizu & Kudoh and *A. kamchatica* subsp. *kawasakiana* (Makino) K. Shimizu & Kudoh, New Combinations. *Acta Phytotaxon Geobotan*, **56**, 163–172.

Shimizu KK, Kudoh H, Kobayashi MJ (2011) Plant sexual reproduction during climate change: Gene function in natura studied by ecological and evolutionary systems biology. *Ann Bot*, **108**, 777–787.

Shimizu-Inatsugi R, Lihová J, Iwanaga H *et al.* (2009) The allopolyploid *Arabidopsis kamchatica* originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. *Mol*

Accepted Article

This article is protected by copyright. All rights reserved.

Ecol, **18**, 4024–4048.

Slotte T, Hazzouri KM, Agren JA *et al.* (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet*, **45**, 831–835.

Smit A, Hubley R, Green P (1996) RepeatMasker Open-3.0. URL http://www.repeatmasker.org.

Sommer D, Delcher A, Salzberg S, Pop M (2007) Minimus: A fast, lightweight genome assembler. *BMC Bioinformatics*, **8**, 64.

Stanke M, Schoffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Tada Silver Mine Historic Site Preservation Association (2007) *Tada ginzan: Shiseki to densho: Ogon densetsu no sato (Tada silver mine: Historic site and lore: Village of golden legend)*. Tada ginzan shiseki hozon kensho kai, Inagawa-cho (in Japanese). Japanese National Bibliography Number 21410864. URL http://www.tadaginzankenshoukai.com/.

Talke IN, Hanikenne M, Krämer U (2006) Zinc-dependent global transcriptional control, transcriptional deregulation, and higher gene copy number for genes in metal homeostasis of the hyperaccumulator *Arabidopsis halleri*. *Plant Physiol*, **142**, 148–167.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Tsuchimatsu T, Kaiser P, Yew C-L, Bachelier JB, Shimizu KK (2012) Recent loss of self-incompatibility by degradation of the male component in allotetraploid *Arabidopsis kamchatica*. *PLoS Genet*, **8**, e1002838.

Tsuchimatsu T, Suwabe K, Shimizu-Inatsugi R *et al.* (2010) Evolution of self-compatibility in *Arabidopsis* by a mutation in the male specificity gene. *Nature*, **464**, 1342–1346.

Van der Auwera GA, Carneiro MO, Hartl C *et al.* (2013) From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. In: *Current protocols in bioinformatics*, pp. 11.10.1–11.10.33. John Wiley & Sons, Inc.

Willems G, Dräger DB, Courbot M *et al.* (2007) The genetic basis of zinc tolerance in the metallophyte *Arabidopsis halleri* ssp. *halleri* (Brassicaceae): An analysis of quantitative trait loci. *Genetics*, **176**, 659–674.

Wolf DE, Steets JA, Houliston GJ, Takebayashi N (2014) Genome size variation and evolution in allotetraploid *Arabidopsis kamchatica* and its parents, *Arabidopsis lyrata* and *Arabidopsis*

This article is protected by copyright. All rights reserved.

Accepted Article

halleri. *AoB Plants*, **6**.

Accepted Article

This article is protected by copyright. All rights reserved.